Leveraging Low-Rank Relations between Surrogate Tasks in Structured Prediction

Giulia Luise [1] Dimitris Stamos [1] Massimiliano Pontil [1,2] Carlo Ciliberto [3]

[1]Dept. of Computer Science, University College of London, UK. [2]CSML, Istituto Italiano di Tecnologia, Genova, Italy. [3]Dept. of Electrical and Electronic Engineering, Imperial College London, UK.

SETTING

Setting: $\mathcal{X}$ input space, $\mathcal{Y}$ set of **structured outputs**, ρ probability distribution on $\mathcal{X} \times \mathcal{Y}$, loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

Goal: learn $f : \mathcal{X} \to \mathcal{Y}$ minimizing the *expected risk*

$$\mathcal{E}(f) := \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y)\, d\rho(x, y),$$

given only n observations $(x_i, y_i)_{i=1}^n \sim \rho$.

BACKGROUND

Surrogate methods are strategies to address supervised learning problems with **structured outputs**.

coding
$\mathsf{c} : \mathcal{Y} \to \mathcal{H}$

decoding
$\mathsf{d} : \mathcal{H} \to \mathcal{Y}$

surrogate problem
$\mathcal{H}$ Hilbert space
$\mathsf{L} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$
$\mathcal{R}(g) = \int_{\mathcal{X}\times\mathcal{Y}} \mathsf{L}(g(x), \mathsf{c}(y))d\rho$
ERM $\to \hat{g}$

The class of **Structure Encoding Loss Functions (SELF)** implicitly define a *coding* function c, since they can be written as

$$\ell(y, y') = \langle \mathsf{c}(y), V\mathsf{c}(y') \rangle_{\mathcal{H}}$$

for some bounded linear operator $V : \mathcal{H} \to \mathcal{H}$.

This leads to a surrogate framework by choosing

Surrogate loss: $\mathsf{L}(h, h') = \|h - h'\|^2$
Decoding: $\mathsf{d}(h) = \operatorname{argmin}_{y\in\mathcal{Y}} \langle \mathsf{c}(y), Vh \rangle_{\mathcal{H}}$ for $h \in \mathcal{H}$

Surrogate estimator $\hat{g}$: LS + Tikhonov reg

$$\min_{g:\mathcal{X}\to\mathcal{H}} n^{-1} \sum_{i=1}^n \|g(x_i) - \mathsf{c}(y_i)\|_{\mathcal{H}}^2 + \lambda\|g\|_{\mathsf{HS}}^2$$

$$\hat{g}(x) = \sum_{i=1}^n \alpha_i(x)\mathsf{c}(y_i), \quad \alpha(x) = (K_x + n\lambda I)^{-1} v_x,$$

Decoded estimator $\hat{f} = \mathsf{d} \circ \hat{g}$ (Loss trick!)

$$\hat{f}(x) = \mathsf{d}(\hat{g}(x)) = \operatorname*{argmin}_{y\in\mathcal{Y}} \sum_{i=1}^n \alpha_i(x)\ell(y, y_i),$$

Comparison Inequality (links *structured* and *surrogate* problem):

$$\mathcal{E}(\mathsf{d} \circ g) - \mathcal{E}(f^*) \le 2\|V\|\sqrt{(\mathcal{R}(g) - \mathcal{R}(g^*))}$$

The framework is practical when $\mathcal{H}$ is infinite dimensional!

CONTRIBUTIONS IN A NUTSHELL

- Study the **interplay** between **surrogate methods** for structured prediction and **multitask learning (MTL)** methods.
- Propose a **trace norm regularization** algorithm that does not require explicit knowledge of the surrogate framework (e.g. coding/decoding).
- Derive **excess risk bounds** for Low-rank Structured Prediction.
- Identify regimes where the proposed MTL estimator exhibits **better generalization performance** than its "independent task" counterpart.

LOW RANK ESTIMATOR

Is it useful to enforce low-rank structures among surrogate tasks?

Can we deal with infinite dimensional surrogate spaces when enforcing such low-rank structures?

$$\min_{g:\mathcal{X}\to\mathcal{H}} n^{-1} \sum_{i=1}^n \|g(x_i) - \mathsf{c}(y_i)\|_{\mathcal{H}}^2 + \lambda\|g\|_* \quad (1)$$

This yields the decoded estimator:

$$\hat{f}(x) = \operatorname*{argmin}_{y\in\mathcal{Y}} \sum_{i=1}^n \alpha_i^{\mathsf{tn}}(x)\, \ell(y, y_i)$$

where $\alpha^{\mathsf{tn}} : X \to \mathbb{R}$ is given by the algorithm below

Algorithm 1 - Low-rank Structured Prediction

Input: $K_x, K_y \in \mathbb{R}^{n\times n}$ empirical kernel matrices for input and output data, λ regularizer, r rank, ν step size, k number of iterations.

Initalize: Sample $M_0, N_0 \in \mathbb{R}^{n\times r}$ randomly.

For $j = 0, \dots, k$:
$M_{j+1} = (1 - \lambda\nu)M_j - \nu(K_x M_j N_j - I)K_y N_j$
$N_{j+1} = (1 - \lambda\nu)N_j - \nu(N_j M_j^\top K_x - I)K_x M_j$

Return: The weighting function $\alpha^{\mathsf{tn}} : \mathcal{X} \to \mathbb{R}^n$ with $\alpha^{\mathsf{tn}}(x) = N_k M_k^\top v_x$ for any $x \in \mathcal{X}$

Let $g(\cdot) = AB^*\phi(\cdot)$ with $\phi : \mathcal{X} \to \mathcal{F}$, $A : \mathcal{H} \to \mathbb{R}^r$, $B : \mathcal{F} \to \mathbb{R}^r$.
Using the **variational formulation** of trace norm, (1) can be rewritten as

$$\min_{A,B,r} n^{-1} \sum_{i=1}^n \|AB^*\phi(x_i) - \psi(y_i)\|_{\mathcal{H}}^2 + \lambda\left(\|A\|_{\mathsf{HS}}^2 + \|B\|_{\mathsf{HS}}^2\right), \quad (2)$$

Theorem 1. (Loss trick) The k-th iterate of GD on problem (2) is $\hat{g}_k(x) = \sum_{i=1}^n \alpha_k^{\mathsf{tn}}(x)\, \mathsf{c}(x)$, with α_k^{tn} obtained after k steps of Alg. 1.

EXCESS RISK BOUNDS

Theorem 2. (Informal) Assume that $\mathcal{Y}$ is compact and that $\|g_*\|_* < +\infty$. Then, the minimizer $\hat{g}$ of (1) satisfies

$$\mathcal{R}(\hat{g}) - \min_{g:\mathcal{X}\to\mathcal{H}} \mathcal{R}(g) \;\le\; \mathsf{M}\, n^{-\frac{1}{2}} \qquad w.h.p.$$

The comparison inequality automatically lifts to the *original* problem any advantage deriving from enforcing low-rankness in the *surrogate* problem.

Corollary 3. The estimator $\hat{f} = \mathsf{d} \circ \hat{g}$ satisfies

$$\mathcal{E}(\hat{f}) - \min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f) \;\le\; \sqrt{\mathsf{M}}\, n^{-\frac{1}{4}} \qquad w.h.p.$$

The **constant M** determines whether the MTL-inspired trace-norm estimator is favorable over the "independent" task learning counterpart:

low-rank: $\mathsf{M} = \|C\|_{op}^{1/2}\|g_*\|_*^2 + \dots$
no low-rank: $\mathsf{M} = (1+\|C\|_{op}^{1/2})\|g_*\|_{\mathsf{HS}}^2 + \dots$

Low-rank is beneficial in the following regime:

- The tasks have a low rank structure $\to \|g_*\|_* \sim \|g_*\|_{\mathsf{HS}}$
- The marginal $\rho_{\mathcal{X}}$ has $\|C\|_{op}^{1/2} \ll 1$

Example. $\mathcal{X} = \mathbb{R}^d$, $\rho_{\mathcal{X}}$ is uniform on the unit sphere: $\|C\|_{op}^{1/2} = 1/\sqrt{d}$.

Remark 1: Our analysis extends to (nonlinear) multitask learning.
Remark 2: The bound in Thm. 2 holds in general for least-squares regression with trace norm regularization.

EXPERIMENTS

We evaluated the empirical performance of the proposed method on ranking applications, specifically the pairwise ranking setting.

	ml100k	jester1	sushi
AdaRank	0.509 (±0.007)	0.534 (±0.009)	0.588 (±0.051)
Random Forests	0.526 (±0.022)	0.548 (±0.001)	0.566 (±0.010)
SVMrank	0.513 (±0.009)	0.507 (±0.007)	0.541 (±0.005)
SELF + $\|\cdot\|_{\mathsf{HS}}$	0.312 (±0.005)	0.386 (±0.005)	0.391 (±0.003)
(Ours) SELF + $\|\cdot\|_*$	**0.156** (±0.005)	**0.247** (±0.002)	**0.313** (±0.003)

REFERENCES

[1] C. Ciliberto et al, *A Consistent Regularization Approach for Structured Prediction*, NIPS 2016
[2] F. Bach, *Consistency of trace norm minimization*, JMLR 2008
[3] Srebro et al, *Maximum-margin matrix factorization*, NIPS 2005